

Examining Trace Files with Jumpshot

Rusty Lusk

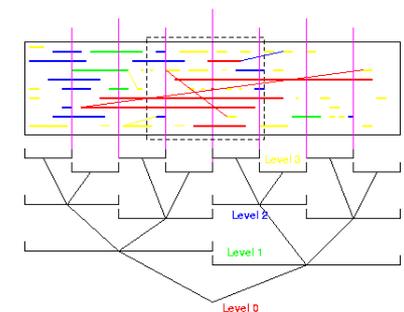
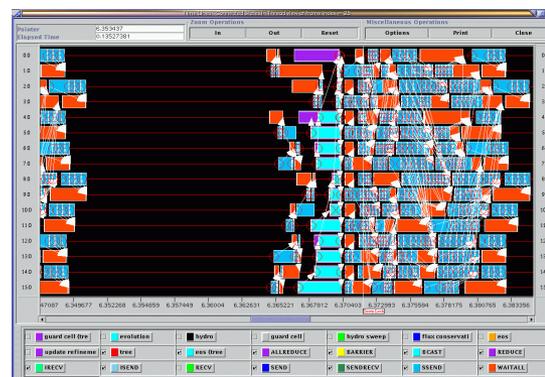
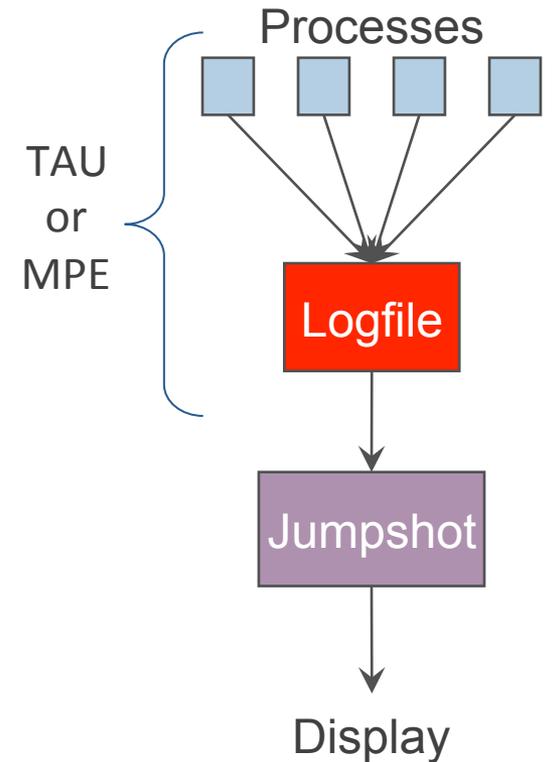
Anthony Chan

Mathematics and Computer Science Division

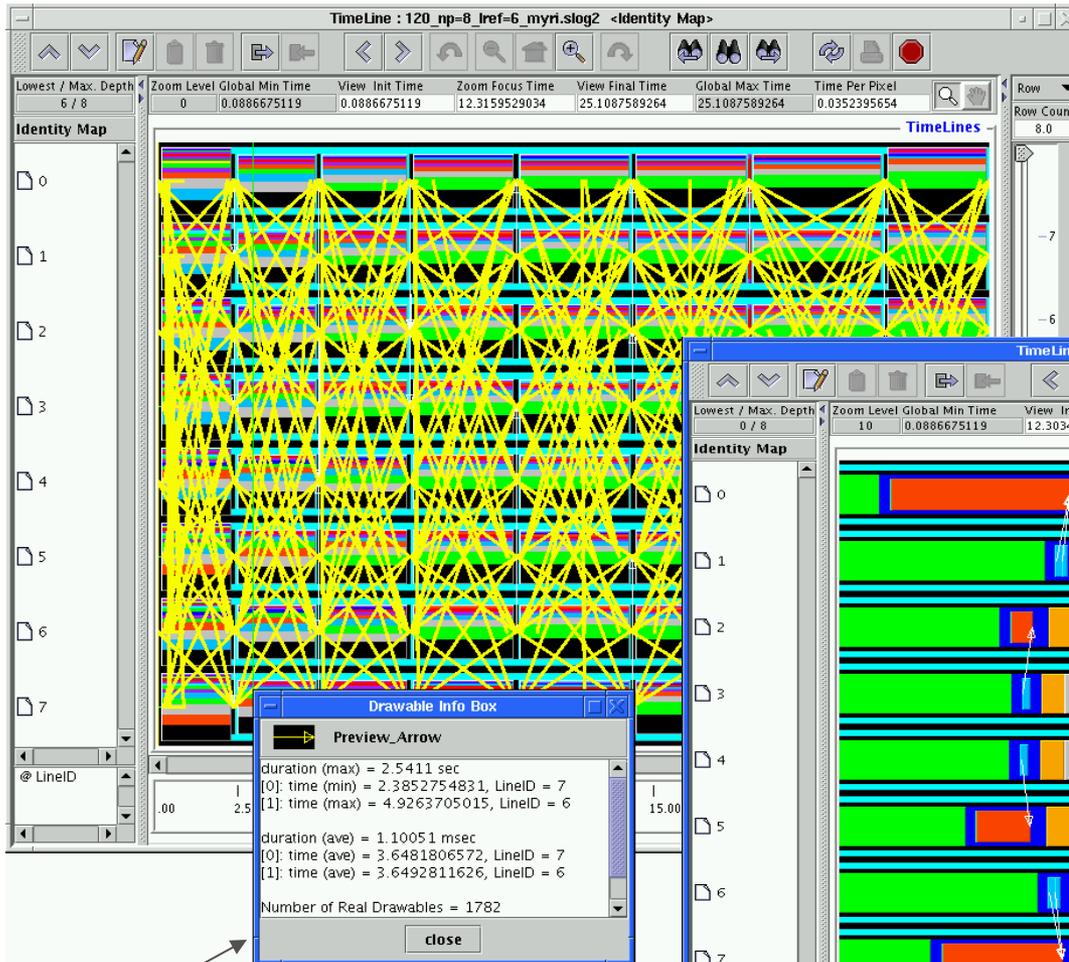
Argonne National Laboratory

Performance Visualization with Jumpshot

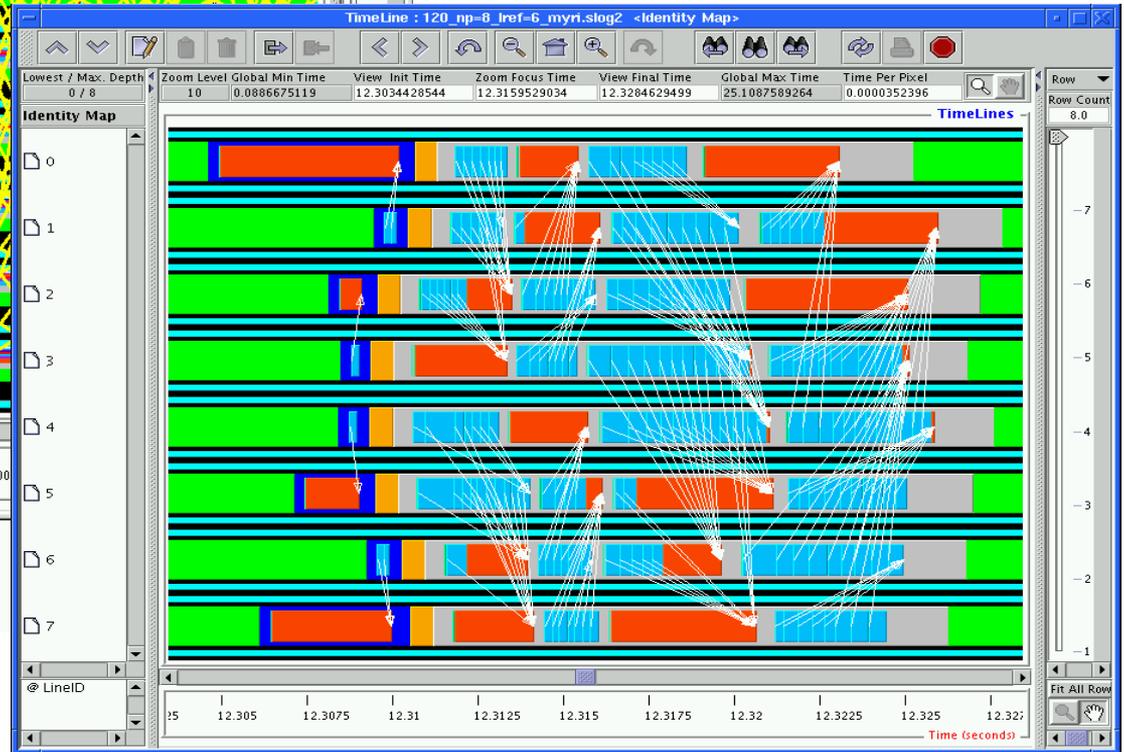
- For detailed analysis of parallel program behavior, timestamped events are collected into a log file during the run.
- A separate display program (Jumpshot) aids the user in conducting a post mortem analysis of program behavior.
- We use an indexed file format (SLOG2) that uses a preview to select a time of interest and quickly display an interval, without ever needing to read much of the whole file.



Viewing Multiple Scales



Detailed view shows opportunities for optimization



Each line represents 1000's of messages

1000x zoom



Pros and Cons of this Approach

■ Cons:

- Scalability limits
 - Screen resolution
 - Big log files, although
 - Jumpshot can read SLOG2 files fast
 - Tracefile generator can be instructed to log a small number of event types
- Use for debugging only indirect

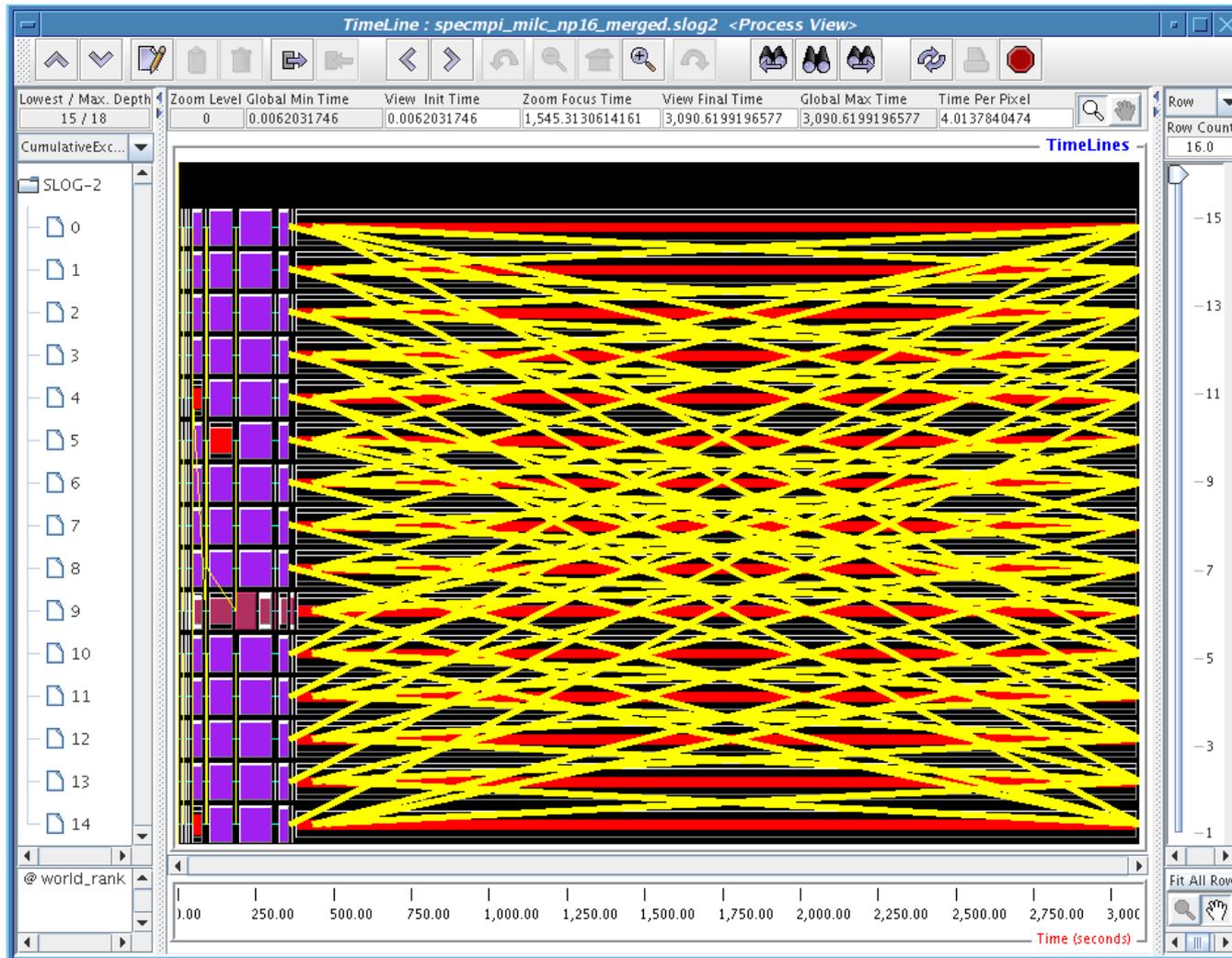
■ Pros:

- Portable, since based on MPI profiling interface
- Provides details other methods might not reveal
- SLOG2 files can be produced either by MPE (part of MPICH) or TAU
- Works with threads (both pthreads and OpenMP) as well as MPI
- Aids understanding of program behavior
 - Almost always see something unexpected



Looking at MILC in SPEC2007

- Curious amount of All_reduce in initialization - why?



MILC

- The answer, and how

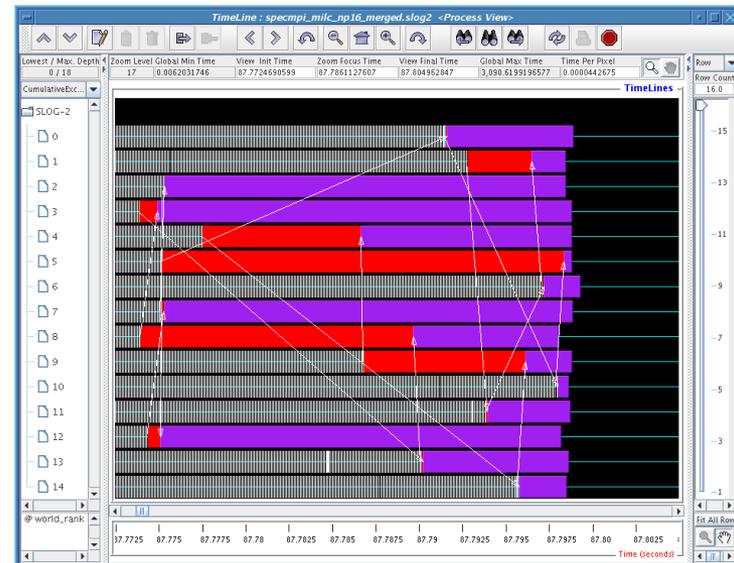
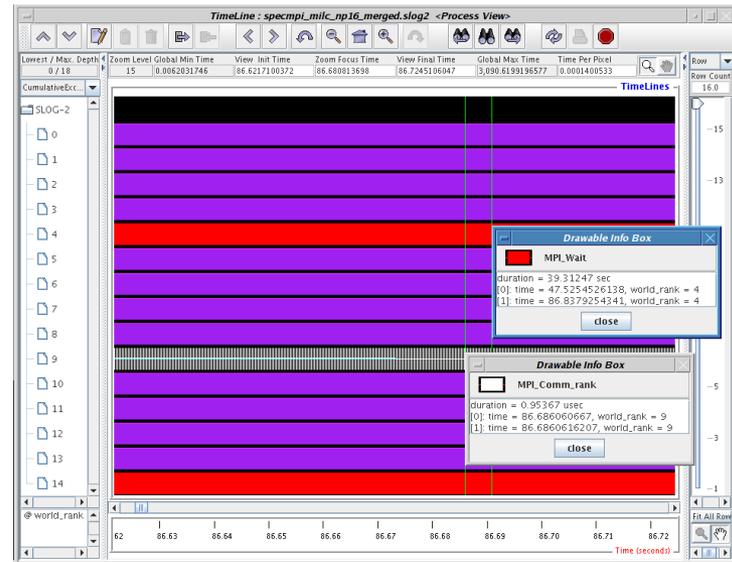
Legend : specmpi_milc_np16_merged.slog2

Topo	Name	S	count	
	Preview_Arrow	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0
	message	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	120330
	Preview_State	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0
	CLOG_Buffer_write2disk	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	2040
	MPE_Irecv_waited	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	120330
	MPI_Allreduce	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	8850
	MPI_Barrier	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	30
	MPI_Bcast	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	60
	MPI_Comm_rank	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	266720746
	MPI_Comm_size	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	24540
	MPI_Comm_split	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	15
	MPI_Irecv	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	120330
	MPI_Isend	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	120330
	MPI_Wait	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	240660
	Preview_Event	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0
	MPE_Comm_finalize	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	15
	MPE_Comm_init	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	15

All

Select Deselect

close



MILC

- The answer - why

- Deep in innermost of quadruply nested loop, an innocent-looking line of code:

```
If ( i > myrank() ) ...
```

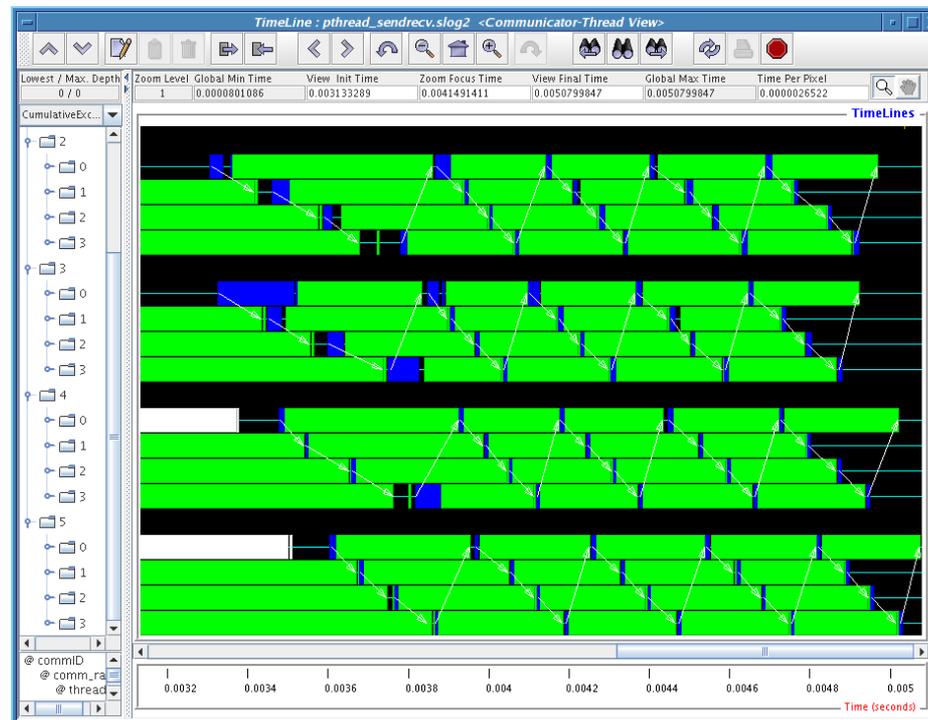
And myrank is a function that calls MPI_Comm_rank (266 million times)

- It actually doesn't cost that much here, but
- It illustrates that you might not know what your code is doing what you think it is
 - Not a scalability issue (found on small # of processes)



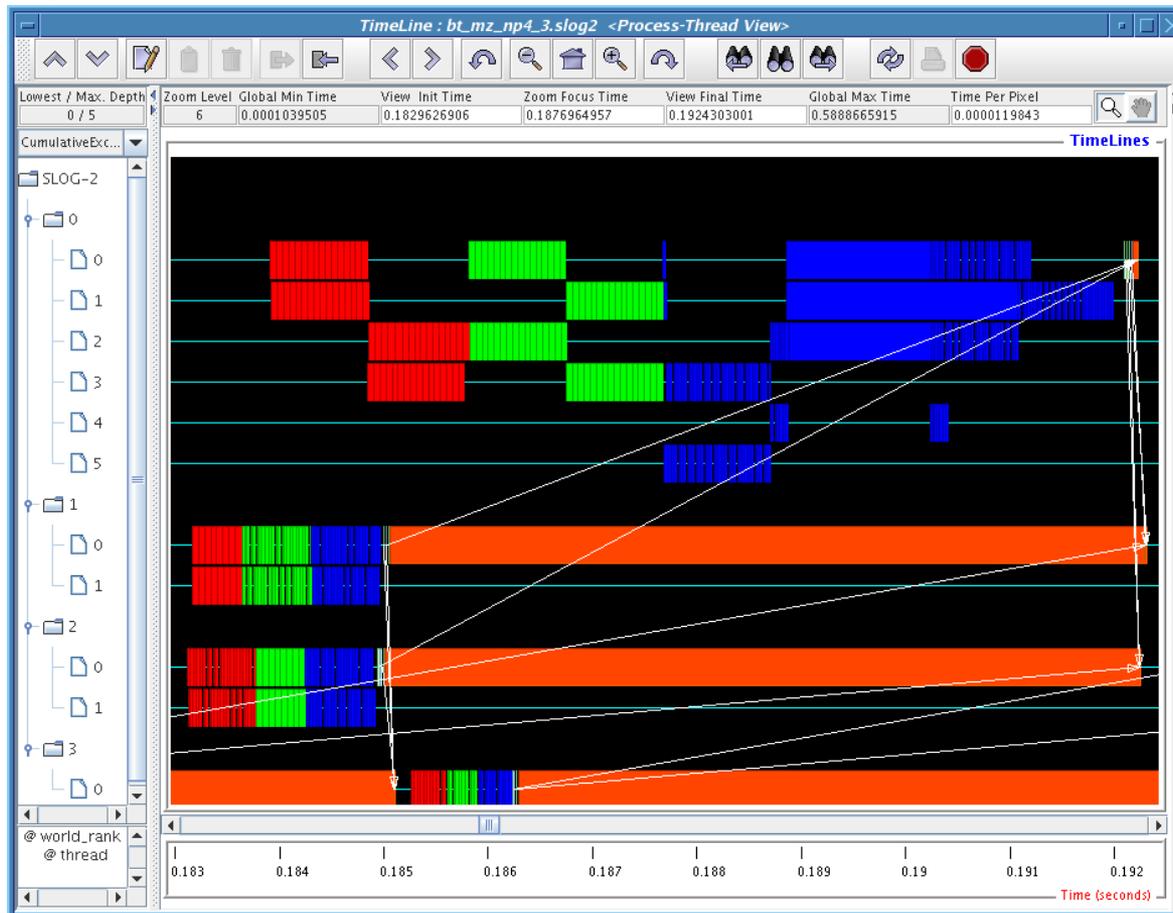
Visualizing Hybrid Programs with Jumpshot

- Recent additions to Jumpshot for multithreaded and hybrid programs that use Pthreads (and most OpenMP implementation use Pthreads) or UPC
 - Separate timelines for each thread id
 - Support for grouping threads by communicator as well as by process



It Might Not Be Doing What You Think

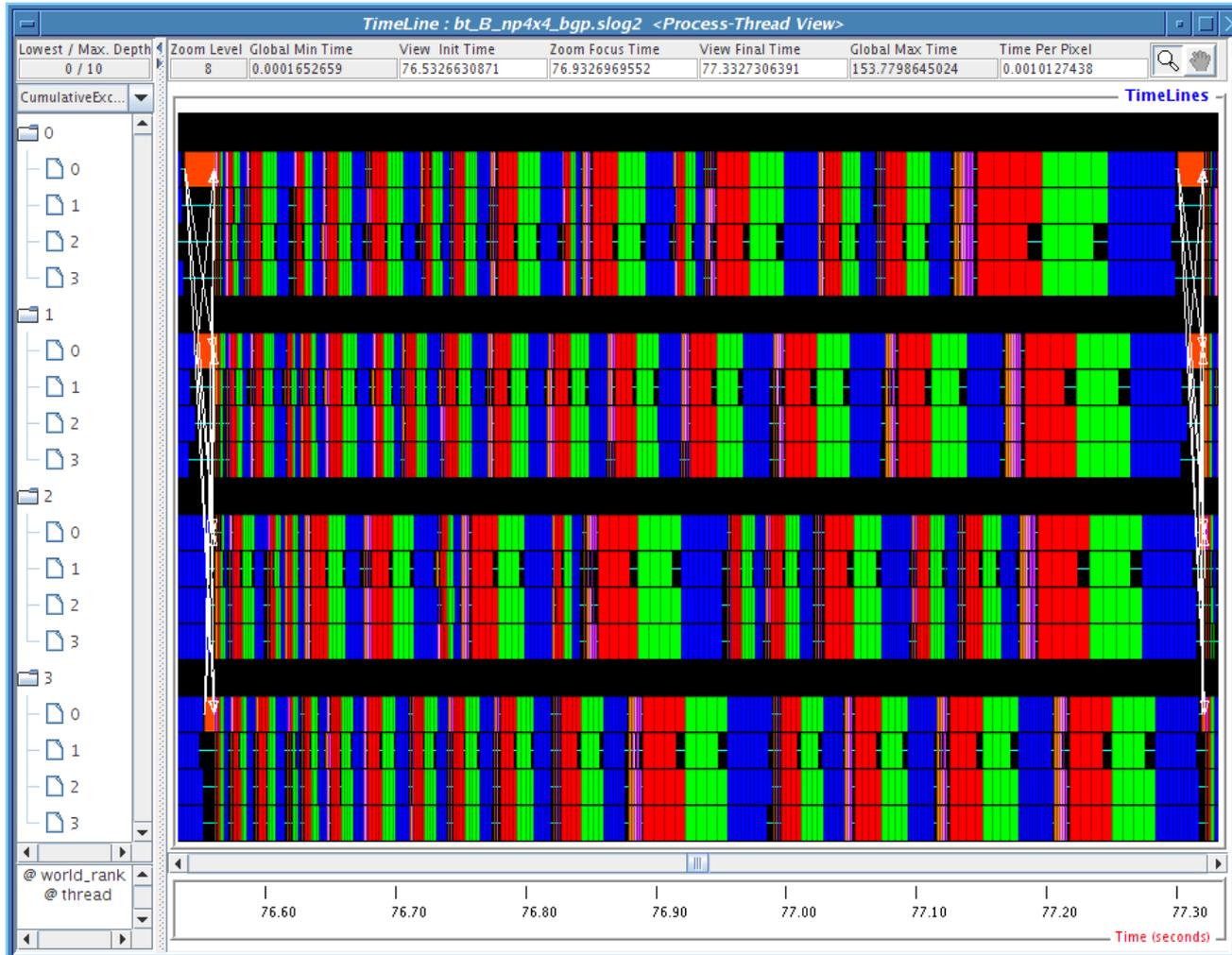
- An early NAS Parallel Benchmark run:



- Nasty interaction between the environment variables OMP_NUM_THREADS and NPB_MAX_THREADS

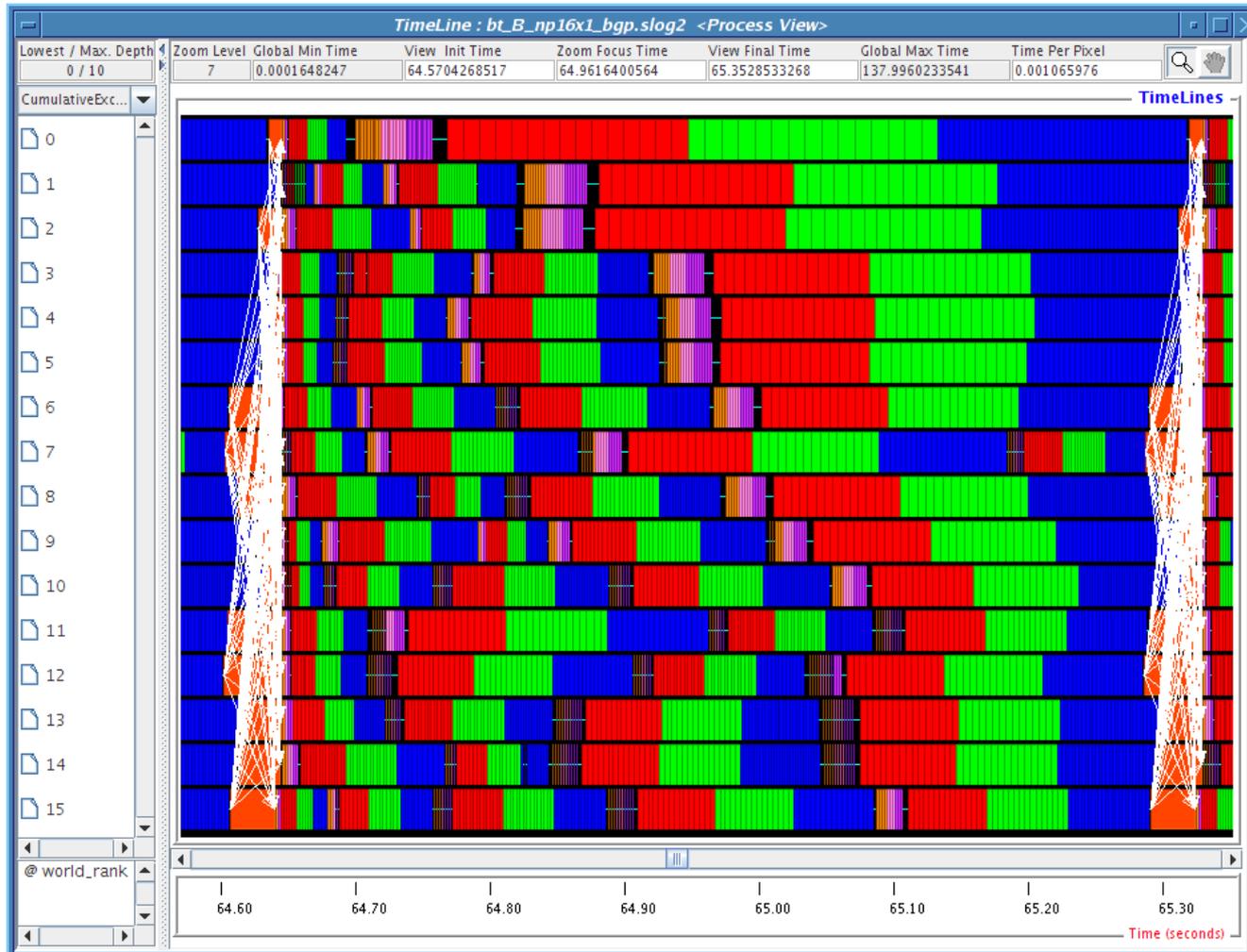
More Like What You Expect

- BT class B on 4 BG/P nodes, using OpenMP on each node



MPI Everywhere

- BT class B on 4 BG/P nodes, using 16 MPI processes



Using Jumpshot with Cobalt

- It is not difficult to turn other types of log files into SLOG2 files
- We thought it would be interesting to look at Intrepid usage both by the year and by the day
- It was easy to parse the logs emitted by Cobalt, the job scheduler used on Intrepid and turn them into SLOG2 files
 - Timelines represent entire racks, which are the scheduling units
 - States represent the duration of a job on a rack
 - State colors represent accounts the jobs are charged to
- One can easily identify idle times, system times, the effectiveness of the backfill scheduling strategy, etc.
- Various approaches used by users can be identified



The Cobalt Logs

11/07/2011 00:23:40;Q;428386;queue=prod-short

11/07/2011 00:31:03;E;428238;Exit_status=0 Resource_List.ncpus=2048 Resource_List.nodect=512 Resource_List.walltime=1:30:00
account=LatticeQCD approx_total_etime=79635 args=-qmp-geom,4,8,8,8,logs/
02256_par1src13sgnPP_P1_D_s01_c012/02256_par1src13sgnPP_P1_D_s01_c012.job.lua,config.lua,main.lua
ctime=1320541718.51 cwd=/intrepid-fs0/users/jrgreen/persistent/prod/RBC_DWFtw/32c64_175_0001/run1/bkw_prop.2s3c
end=1320625863.28 etime=1320541718.51 exe=/intrepid-fs0/users/jrgreen/persistent/prod/RBC_DWFtw/32c64_175_0001/run1/
bkw_prop.2s3c/bin/qlua-bkend.new exec_host=ANL-R21-M1-512 group=unknown jobname=02256_par1src13sgnPP_P1_D_s01_c012
mode=vn priority_core_hours=7990107 qtime=1320541718.51 queue=backfill resources_used.location=ANL-R21-M1-512
resources_used.nodect=512 resources_used.walltime=1:15:08 session=unknown start=1320621354.92 user=jrgreen

11/07/2011 00:31:11;S;428386;Resource_List.ncpus=512 Resource_List.nodect=512 Resource_List.walltime=1:00:00
account=MADNESS_MPQC_esp args= ctime=1320625420.26 cwd=/gpfs/home/robert/trunk/src/apps/moldft etime=1320625420.26
exe=/gpfs/home/robert/trunk/src/apps/moldft/./moldft exec_host=ANL-R21-M1-512 group=unknown jobname=N/A mode=smp
qtime=1320625420.26 queue=prod-short session=unknown start=1320625871.89 user=robert

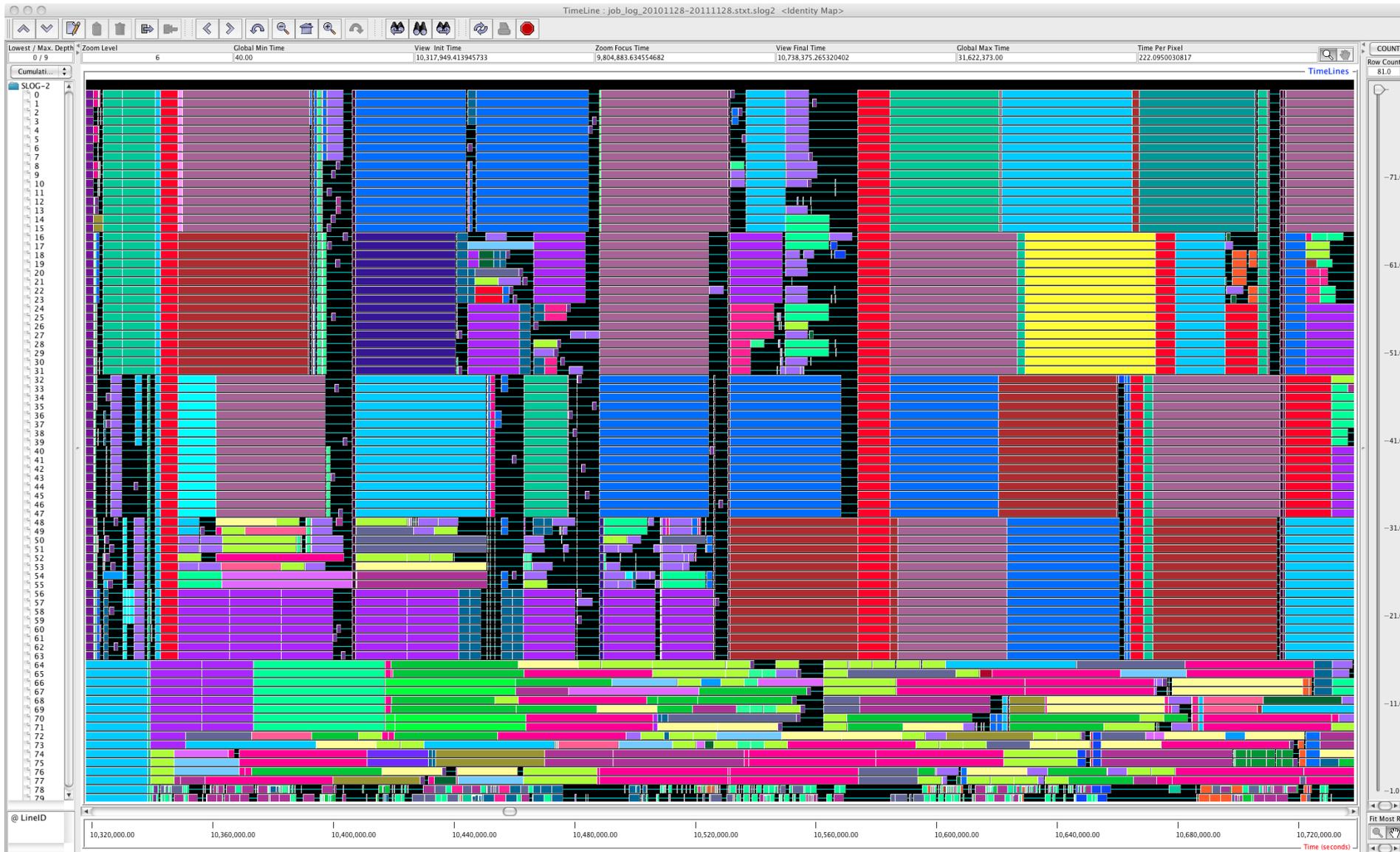
11/07/2011 00:31:13;Q;428387;queue=backfill11/07/2011 00:31:56;E;428189;Exit_status=0 Resource_List.ncpus=2048
Resource_List.nodect=512 Resource_List.walltime=1:30:00 account=LatticeQCD approx_total_etime=115758 args=-qmp-geom,
4,8,8,8,logs/02248_par0src13sgnPM_P1_D_s01_c012/02248_par0src13sgnPM_P1_D_s01_c012.job.lua,config.lua,main.lua
ctime=1320505614.17 cwd=/intrepid-fs0/users/jrgreen/persistent/prod/RBC_DWFtw/32c64_175_0001/run1/bkw_prop.2s3c
end=1320625916.77 etime=1320505614.17 exe=/intrepid-fs0/users/jrgreen/persistent/prod/RBC_DWFtw/32c64_175_0001/run1/
bkw_prop.2s3c/bin/qlua-bkend.new exec_host=ANL-R24-M1-512 group=unknown jobname=02248_par0src13sgnPM_P1_D_s01_c012
mode=vn priority_core_hours=7704134 qtime=1320505614.17 queue=backfill resources_used.location=ANL-R24-M1-512
resources_used.nodect=512 resources_used.walltime=1:15:46 session=unknown start=1320621369.98 user=jrgreen

11/07/2011 00:32:11;Q;428388;queue=backfill

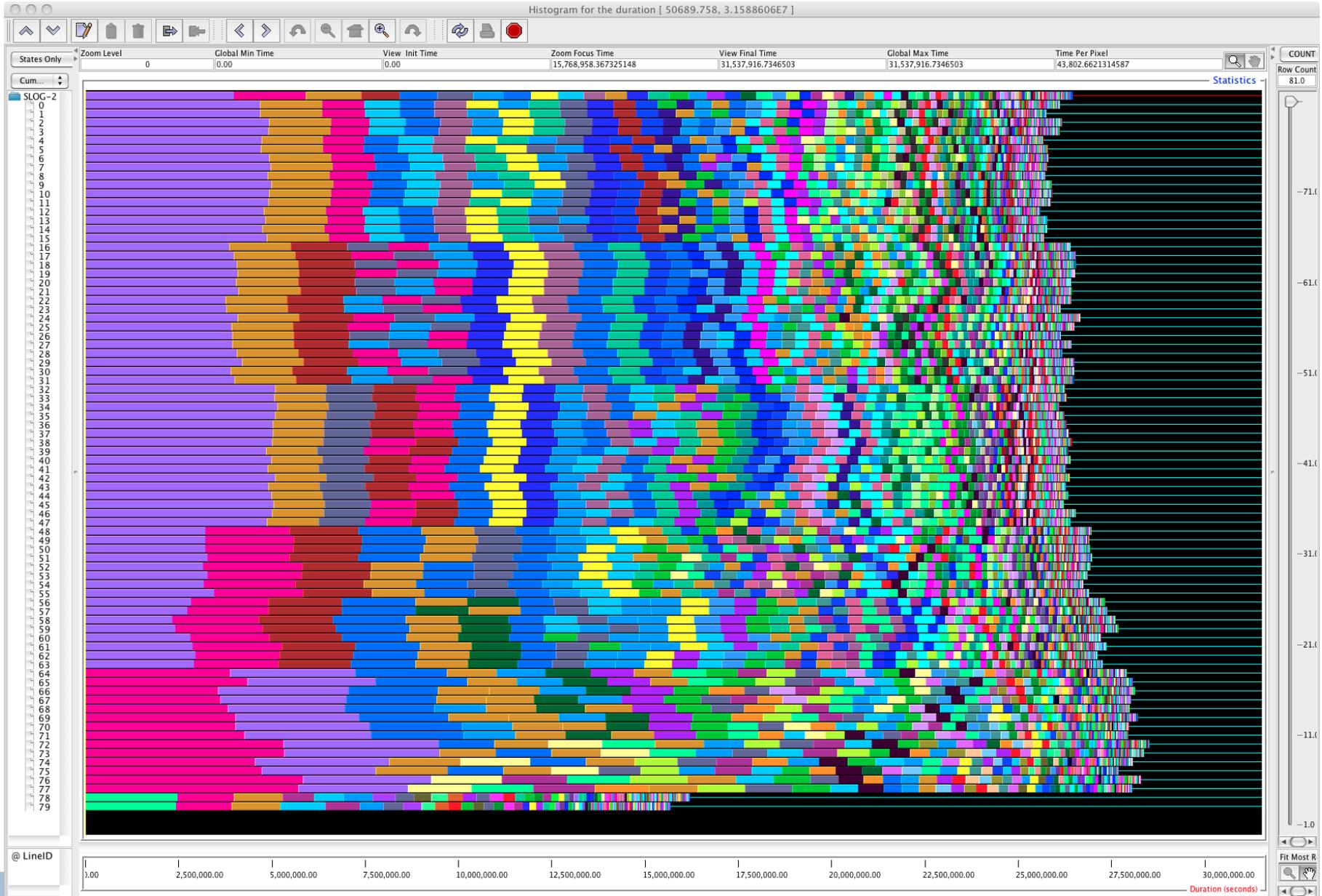
11/07/2011 00:32:33;S;428237;Resource_List.ncpus=2048 Resource_List.nodect=512 Resource_List.walltime=1:30:00 account=LatticeQCD
args=-qmp-geom,4,8,8,8,logs/
02256_par1src13sgnPP_P1_U_s01_c012/02256_par1src13sgnPP_P1_U_s01_c012.job.lua,config.lua,main.lua
ctime=1320541714.79 cwd=/intrepid-fs0/users/jrgreen/persistent/prod/RBC_DWFtw/32c64_175_0001/run1/bkw_prop.2s3c
etime=1320541714.79 exe=/intrepid-fs0/users/jrgreen/persistent/prod/RBC_DWFtw/32c64_175_0001/run1/bkw_prop.2s3c/bin/qlua-
bkend.new exec_host=ANL-R24-M1-512 group=unknown jobname=02256_par1src13sgnPP_P1_U_s01_c012 mode=vn
qtime=1320541714.79 queue=backfill session=unknown start=1320625953.14 user=jrgreen



Jumpshot view of Cobalt logs



Statistics View



Conclusions

- Jumpshot can be useful in *understanding* the behavior of MPI, OpenMP, and hybrid programs.
 - Sometimes you need to see the microscopic details
- Best used at relatively small scale (< 300 time lines), but that is where a significant number of performance problems can be studied.
- Other tools available as you scale up, once performance on a small number of nodes has been understood and tuned.



